REDUCING HALLUCINATIONS IN GENERATIVE MODELS THROUGH TRUNCATED STATISTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucinations—where generative models produce invalid or nonsensical outputs—remain a critical challenge for reliable deployment. We present the first computationally and query-efficient algorithm that provably addresses the hallucination problem by actively querying the model's own invalid outputs. Specifically, we control the hallucination rate up to an arbitrarily small multiplicative slack while approximately maximizing the likelihood of valid target examples via projected stochastic gradient descent. Our method works in very general settings with arbitrary validity sets and valid observations, within fixed powerful exponential families satisfying stated regularity and sampling assumptions. Our approach is enabled by a novel connection to the field of truncated statistics and settles an open problem posed by Hanneke et al. (2018).

1 INTRODUCTION

Generative models can produce outputs that violate factual, semantic, or application-specific constraints. Such hallucinations limit their use in settings that require reliable outputs (Ji et al., 2023; Kalai & Vempala, 2024). We adopt a distribution learning formulation. Let X be the output space, $V \subseteq X$ its measurable valid subset, and $S = X \setminus V$ the invalid subset. We observe valid examples $x_{1:n} \subseteq V$ and can query the invalidity indicator $\mathbf{1}_S$. For a parametric family $\{p_\theta : \theta \in \Theta\}$, our objective is

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) \quad \text{subject to} \quad p_\theta(S) \leq \alpha, \qquad \mathcal{L}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i). \tag{1}$$

Thus likelihood measures fit to the examples, while $p_\theta(S)$ measures the hallucination rate. These are empirical optimization guarantees; population generalization requires a separate statistical argument.

A mixture with a fixed valid answer can make invalid mass arbitrarily small. However, assigning most of the probability to that answer can substantially reduce likelihood on the observed data. The constraint in (1) asks for the best likelihood attainable at a prescribed hallucination rate, rather than validity alone.

Why the constraint is non-convex. Even a one-dimensional Gaussian shows the difficulty. Take $p_\mu = \mathcal{N}(\mu, 1)$ and $S = (-1, 1)$. Writing Φ and ϕ for the standard normal distribution and density,

$$p_\mu(S) = \Phi(1 - \mu) - \Phi(-1 - \mu), \qquad p_0(S) = 2\Phi(1) - 1 \approx 0.6827.$$

This mass is even in μ, tends to zero as $|\mu| \to \infty$, and is strictly decreasing for $\mu > 0$, since its derivative is $\phi(1 + \mu) - \phi(1 - \mu) < 0$. Thus, for $0 < \alpha < p_0(S)$, the feasible means form two disjoint rays. Large positive and negative means are feasible, but their midpoint is not. The Gaussian negative log-likelihood is convex in μ; the invalid-mass constraint causes the non-convexity. Our method therefore searches through convex likelihood sublevels instead of projecting onto the feasible set.

Passive observations do not suffice for general valid distribution learning. Hanneke et al. (2018) introduced active distribution learning with an invalidity oracle. They established a worst-case query barrier for proper learning, gave statistically and query-efficient improper procedures, and posed the question of computational efficiency under relaxed validity. We resolve this computational question

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in the affirmative for the fixed powerful exponential-family model specified below. The validity set can be arbitrary, and the valid data need not be generated by a member of the family. Without a capacity assumption, even finding a feasible Gaussian model can be hard.

Hardness and the modeling requirement.

Claim 1 (Hardness without a capacity assumption)**.** *For $p_\theta = \mathcal{N}(\theta, I_d)$ and an efficiently computable invalidity indicator, deciding whether a parameter with $p_\theta(S) \le 1/4$ exists is NP-hard. The instance can include a specified valid observation.*

The reduction from Max-Cut in Appendix G establishes a constant feasibility gap. This motivates a capacity requirement on the initial family.

Definition 2 (Powerfulness)**.** Fix S and valid observations $x_{1:n}$. A family $\mathcal{P}$ is α-powerful if, for every $p \in \mathcal{P}$ with $p(S) > \alpha$, there is $p' \in \mathcal{P}$ with $p'(S) = \alpha$ and

$$\frac{p'(S)}{p(S)} \le \exp\left(\frac{1}{n}\sum_{i=1}^{n}[\log p'(x_i) - \log p(x_i)]\right) = e^{\mathcal{L}(p)-\mathcal{L}(p')}. \tag{2}$$

The family is powerful if this holds for every $\alpha \in (0, 1)$. All likelihoods in this definition are finite. A family satisfying the condition for every valid data set is powerful uniformly over such data sets.

The right-hand side is the ratio of geometric-mean likelihoods. Thus reducing invalid mass by a factor r permits an increase of at most $\log(1/r)$ in average negative log-likelihood (NLL). Powerfulness is the capacity requirement used by our algorithm; it does not assert that every exponential family already has this property.

Our result. For a fixed powerful exponential family with well-conditioned sufficient statistics, uniform sub-exponential tails, and the sampling and convex-optimization access stated in Theorem 8, we obtain an approximately optimal solution of (1). Its invalid mass is at most $(1 + \varepsilon)\alpha$, its empirical NLL is at most the constrained optimum plus ε, and the conditional KL divergence from a constrained optimum to the returned model is at most ε. The number of truncated-sampling calls is polynomial in dimension, inverse accuracy, the regularity parameters, and $\log(1/\alpha)$. Ordinary sampling and invalidity queries used for mass estimation introduce a polynomial dependence on $1/\alpha$.

The family is fixed before the algorithm starts. To explain why powerfulness is a natural expressivity requirement, Proposition 12 constructs an augmented exponential family that can transfer excess invalid mass to a valid atom while preserving densities at existing valid points. The original family belongs to its closure. This construction establishes capacity; it does not establish the covariance or sampling assumptions of the optimization theorem. If augmentation is used to select a model, those assumptions must be checked for the selected family before optimization.

The mass transfer is explicit. Given $p(S) = s > \alpha$, add a designated valid output x^* and define

$$q(x) = p(x)\ (x \in V), \qquad q(x) = \frac{\alpha}{s}p(x)\ (x \in S), \qquad q(x^*) = s - \alpha.$$

The total mass is one, all original valid densities are preserved, and $q(S) = \alpha$. In an exponential family this operation can be represented by adding the statistics $\mathbf{1}_S$ and $\mathbf{1}_{\{x^*\}}$. The same transfer works from every member of that augmented family, including members already assigning positive mass to x^*. This explains the capacity condition by a concrete increase in expressivity. The optimization theorem applies only after the family has been chosen and its other assumptions checked; the algorithm does not add statistics during optimization.

Token-level softmax distributions are categorical exponential families, and exponential-family approximation results cover rich classes of distributions (Bengio et al., 2003; Vaswani et al., 2017; Barron & Sheu, 1991; Sriperumbudur et al., 2017). Our convex analysis concerns natural parameters of the fixed family. It does not imply convexity in the weights of an arbitrary neural network or curved parametrization.

An initial family satisfying the capacity condition. Powerfulness can also hold without augmentation. Consider $p_\mu = \mathcal{N}(\mu, 1)$ with $V = [0, \infty)$ and $S = (-\infty, 0)$. For any valid data set, its sample

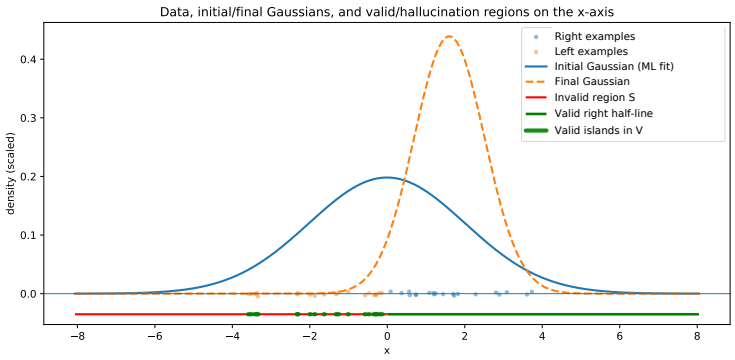


Figure 1: Gaussian illustration, representative run. Blue and orange points are valid observations. The solid curve is the ordinary maximum-likelihood fit; the dashed curve is the final fit. Green segments indicate the valid region V and red indicates the invalid region S.

mean $\overline{x}$ is nonnegative and

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\left[\mathcal{L}(\mu) + \log p_\mu(S)\right] = \mu - \overline{x} - \frac{\phi(\mu)}{\Phi(-\mu)} = \mathbb{E}_{p_\mu}[X \mid X < 0] - \overline{x} < 0.$$

The conditional-mean identity follows by integrating the Gaussian density on the negative half-line. If $p_\mu(S) > \alpha$, the continuity and strict decrease of $\Phi(-\mu)$ give a unique $\mu' > \mu$ with $p_{\mu'}(S) = \alpha$. Integrating the displayed derivative yields

$$\mathcal{L}(\mu') - \mathcal{L}(\mu) \le \log \frac{p_\mu(S)}{\alpha},$$

which is exactly powerfulness. It holds uniformly over all valid data sets. Here the sufficient statistic is $T(x) = x$, its covariance is one, and its centered tails are Gaussian with constants independent of μ. The NLL sublevels are the intervals $[\overline{x} - \sqrt{2c}, \overline{x} + \sqrt{2c}]$, so projection is explicit. Conditional sampling can be implemented by rejection. Thus capacity, curvature, and the required computational access can all hold in the initial family. The earlier non-convex example and the hardness result use different validity sets; the capacity condition depends on both the family and the validity problem.

Numerical illustration. Figure 1 illustrates the likelihood–validity tradeoff in a one-dimensional Gaussian experiment with 40 valid observations. The valid set consists of the nonnegative half-line and small intervals around the negative observations; the remaining outputs are invalid. Ordinary maximum likelihood fits both groups but places substantial mass between the valid regions. The constrained update reduces this invalid mass while controlling the loss of likelihood. Across ten random seeds, the hallucination rate decreases from 0.5239 ± 0.0151 to 0.0338 ± 0.0038, below the target $\alpha = 0.05$ in every run; the errors are standard errors of the mean. This experiment illustrates the objective, rather than verifying every assumption of the theorem. Appendix H gives the setup and results across seeds.

1.1 Technical approach

Write $p_\theta(x) = h(x)e^{\theta^\top T(x) - A(\theta)}$ and $\widehat{\mu} = n^{-1}\sum_i T(x_i)$. The key objective is

$$F_S(\theta) = \mathcal{L}(\theta) + \log p_\theta(S) = A_S(\theta) - \theta^\top \widehat{\mu} + C, \qquad A_S(\theta) = \log \int_S h(x)e^{\theta^\top T(x)}\,\mathrm{d}x. \tag{3}$$

Here $C = -n^{-1}\sum_i \log h(x_i)$ is independent of θ. The Hessian of F_S is the covariance of T under p_θ^S, so F_S is convex. Although its derivatives are those of a truncated-statistics objective, it is not the conditional NLL of the valid observations under p_θ^S: those observations lie outside S.

The main optimization hurdle is to find an approximate minimizer near the boundary of a likelihood sublevel set. A small objective gap alone does not guarantee proximity to this boundary. Powerfulness

gives a common non-increasing direction for F_S on bounded regions. We develop the proxy convex function

$$F_{\rho,v}(\theta) = F_S(\theta) - \rho v^\top(\theta - \theta_0), \tag{4}$$

where θ_0 minimizes $\mathcal{L}$. With constant probability, a random unit vector v has sufficient alignment with that direction. Minimizing the proxy to sufficient accuracy gives a point near the boundary with a nearly optimal F_S value.

We minimize this proxy over nested NLL sublevel sets, following the optimization framework of Karatapanis et al. (2025). Their Proposition C.1 bounds progress across nested sublevel sets, and their Lemmas 3.5–3.7 motivate the truncated moment and covariance estimates. We state the tail assumptions and give the arguments needed for the present empirical-data model. We add the powerfulness construction, the common-direction and boundary arguments, and a conditional KL bound from approximate sublevel optimization. The outer search checks the invalid mass directly; a small successive objective gap alone does not guarantee validity.

1.2 Related work

Hanneke et al. (2018) formalized valid generative modeling and its worst-case barriers. Rittler & Chaudhuri (2024) extend that work in *Distribution Learning with Valid Outputs Beyond the Worst-Case*. Their results include a proper, realizable finite-class guarantee and algorithms under assumptions on the validity class, candidate valid mass, and loss. Some constructions use smoothing or restriction and need not remain in the original family. These results address statistical and query guarantees under different structural assumptions. Our result concerns efficient empirical optimization within a fixed powerful exponential family, with the oracle costs distinguished explicitly.

Truncated statistics studies estimation from observations retained by a selection set. Algorithms for Gaussian and exponential-family models develop the sampling and optimization tools underlying this connection (Daskalakis et al., 2018; 2020; Lee et al., 2023; Karatapanis et al., 2025). In our setting the observations are valid, but the stochastic gradients sample the model conditioned on being invalid. This difference requires using $\widehat{\mu}$ from the observed data rather than assuming observations from an unknown model conditioned on S. Active querying also appears in classical learning theory (Angluin, 1987; Jackson, 1997); here queries describe output validity rather than a prediction label.

2 Model and oracle access

Let $\Theta \subseteq \mathbb{R}^k$ be an open convex natural-parameter domain on which A is finite and twice differentiable. The sufficient statistic is $T : X \to \mathbb{R}^k$. The family has common support, and T is finite almost everywhere. We use

$$\nabla A(\theta) = \mathbb{E}_{p_\theta} T, \qquad \nabla^2 A(\theta) = \mathrm{Cov}_{p_\theta}(T), \qquad p_\theta^S(x) = \frac{p_\theta(x)\mathbf{1}_S(x)}{p_\theta(S)}.$$

The valid set is always V and the invalid set is always $S = X \setminus V$; the oracle $\mathbf{1}_S$ returns one on invalid outputs. We assume $0 < p_\theta(S) < 1$ whenever a conditional distribution is used. Fix valid data with finite likelihood, and set

$$\theta_\alpha^\star \in \arg\min_{\theta:\, p_\theta(S)\le\alpha} \mathcal{L}(\theta), \quad \theta_0 \in \arg\min_{\Theta} \mathcal{L}(\theta), \quad \mathcal{L}_0 = \mathcal{L}(\theta_0), \quad C_c = \{\theta : \mathcal{L}(\theta) \le \mathcal{L}_0 + c\}. \tag{5}$$

The empirical mean $\widehat{\mu}$ is a mean parameter, not θ_0; an interior minimum satisfies $\nabla A(\theta_0) = \widehat{\mu}$. Appendix I derives the log-partition identities, convexity, and the conditional KL formula used below.

Definition 3 (Sample oracles; Karatapanis et al., 2025, Definition 1.3)**.** An ordinary sample oracle returns $Z \sim p_\theta$. A β-truncated sample oracle Sample_S returns $Z \sim p_\theta^S$ when $p_\theta(S) \ge \beta$; its behavior below that threshold is unspecified. An $\Omega(\alpha)$-truncated sample oracle means one with threshold $c_0\alpha$ for a fixed absolute constant $c_0 > 0$ supported by the algorithm's mass lower bound.

We use this $\Omega(\alpha)$ convention in the statements below, with $c_0 = 1/e$ as supplied by the analysis. Rejection sampling implements a truncated call using $1/p_\theta(S)$ ordinary samples and invalidity queries in expectation, hence $O(1/\alpha)$ on the region we use. A truncated-oracle count is not a count of elementary membership queries.

Definition 4 (Uniform sub-exponential statistics)**.** The centered sufficient statistics have uniform sub-exponential scale K on a parameter region if, for every unit u, every parameter in that region, and every $t \geq 0$,

$$\Pr_{p_\theta}\left(\left|u^\top (T - \mathbb{E}_{p_\theta} T)\right| > t\right) \leq 2e^{-t/K}. \tag{6}$$

This is a tail assumption on centered statistics, not just an upper bound on their variance at one parameter.

3 Truncated estimates and nested optimization

The following bounds make the dependence on sub-exponential tails explicit. They are the moment estimates used in truncated optimization (Karatapanis et al., 2025, Lemmas 3.5–3.7), stated with a uniform tail scale.

Lemma 5 (Truncated sufficient statistics)**.** *Suppose (6) holds at θ, let $s = p_\theta(S) > 0$, and put $\ell_s = 1 + \log(1/s)$. For an absolute constant C,*

$$\|\mathbb{E}_{p_\theta^S} T - \mathbb{E}_{p_\theta} T\|_2 \leq CK\ell_s, \qquad \mathrm{Cov}_{p_\theta^S}(T) \preceq CK^2\ell_s^2 I. \tag{7}$$

Every centered unit projection under p_θ^S is sub-exponential with scale at most $CK\ell_s$. For $t > 0$ and $0 < \delta < 1$, the sufficient-statistic mean $\widehat{m} = N^{-1}\sum_i T(Z_i)$ of independent samples $Z_i \sim p_\theta^S$ satisfies $\|\widehat{m} - \mathbb{E}_{p_\theta^S} T\|_2 \leq t$ with probability at least $1 - \delta$ if

$$N \geq C\left(\frac{kK^2\ell_s^2}{t^2} + \frac{\sqrt{k}K\ell_s}{t}\right)\log\frac{2k}{\delta}.$$

The proof in Appendix B derives the conditional moments and the scalar concentration inequality directly. If the uniform scale satisfies $K^2 = O(L)$, the conditional covariance bound is $O(L(1 + \log^2(1/s)))I$. We retain K explicitly when this additional relation is not assumed.

Define the value function $m(c) = \min_{\theta \in C_c} F_S(\theta)$. Whenever the sublevels are compact, these minima exist. A boundary minimizer b_c additionally satisfies $\mathcal{L}(b_c) = \mathcal{L}_0 + c$. For such a minimizer define $r(c) = p_{b_c}(S)$; its value is independent of the choice of boundary minimizer because

$$\log r(c) = m(c) - \mathcal{L}_0 - c. \tag{8}$$

Observation 6 (Mass on a sublevel)**.** *For every $\theta \in C_c$, $p_\theta(S) \geq r(c)e^{\mathcal{L}(b_c) - \mathcal{L}(\theta)} \geq r(c)$. Also, for $d > c$, $r(d) \leq r(c)e^{-(d-c)}$.*

These statements follow by comparing F_S values and using the inclusion $C_c \subseteq C_d$. In particular, they use a boundary minimizer; an arbitrary interior approximate minimizer cannot be substituted without a slack term.

Proposition 7 (Progress along nested sublevels)**.** *If C_H is compact for some $H > 0$, then m is convex and non-increasing on $[0, H]$. For $0 \leq c < d \leq e \leq H$,*

$$m(c) - m(d) \geq \frac{d - c}{e - c}\,[m(c) - m(e)]. \tag{9}$$

Thus on a grid of spacing g up to height H, a successive gap controls later improvement by a factor at most H/g.

This is the progress bound of Karatapanis et al. (2025, Proposition C.1). A short proof is to interpolate minimizers in C_c and C_e: convexity of both $\mathcal{L}$ and F_S gives the defining inequality for convexity of m, and rearranging gives (9). Appendix J gives the full interpolation argument and the resulting bound for inexact objective comparisons.

4 Optimization in the fixed family

Theorem 8 (Generation with a controlled hallucination rate)**.** *Fix $0 < \alpha < 1$, $0 < \varepsilon \leq 1$, $0 < \delta < 1/2$, and valid observations $x_{1:n}$. Let the* initial, fixed *exponential family be powerful for these observations and satisfy*

$$\lambda I \preceq \mathrm{Cov}_{p_\theta}(T) \preceq LI \qquad (\theta \in \Theta), \qquad \lambda > 0.$$

Assume an interior NLL minimizer θ_0 exists, C_H is a compact subset of Θ for $H = 2 + \log(1/\alpha)$, and the centered statistics are uniformly sub-exponential with scale K on C_H. Assume θ_0 is supplied or efficiently computable, together with efficient evaluation of T and $\mathcal{L}$, projection onto C_c to the requested accuracy, ordinary sample access, invalidity queries, and an $\Omega(\alpha)$-truncated sample oracle over S.

There is a randomized algorithm returning $\widehat{\theta}$ such that, with probability at least $1 - \delta$,

$$\mathcal{L}(\widehat{\theta}) \le \mathcal{L}(\theta^\star_\alpha) + \varepsilon, \qquad p_{\widehat{\theta}}(S) \le (1+\varepsilon)\alpha, \qquad \mathrm{KL}(p^S_{\theta^\star_\alpha} \| p^S_{\widehat{\theta}}) \le \varepsilon. \tag{10}$$

After forming $\widehat{\mu}$, truncated-sampling calls and convex-optimization operations are polynomial in $k, L, \lambda^{-1}, K, \varepsilon^{-1}, \log(1/\alpha), \log(1/\delta)$. Ordinary samples and invalidity queries, including mass estimation, are polynomial in the same parameters and $1/\alpha$. Runtime is polynomial in these quantities and the stated evaluation, projection, and sampling costs. The guarantees are conditional on the observed valid data and do not assume that their source distribution belongs to the family.

What the assumptions provide. The covariance bounds control the curvature of the ordinary NLL. Integrating its Hessian along the segment from θ_0 to θ gives

$$\frac{\lambda}{2}\|\theta - \theta_0\|_2^2 \le \mathcal{L}(\theta) - \mathcal{L}_0 \le \frac{L}{2}\|\theta - \theta_0\|_2^2, \qquad \|\nabla\mathcal{L}(\theta)\|_2 \le L\|\theta - \theta_0\|_2.$$

In particular, the search sublevels have bounded diameter, and an objective tolerance controls displacement toward their boundaries. The sub-exponential assumption has a different role: after conditioning on S, it bounds the gradient variance and the smoothness of F_S in terms of the surviving mass. These statistical bounds do not supply the computational oracles. Evaluation of $\mathcal{L}$, projection onto its sublevels, and sampling are separate access assumptions, with their costs included in the theorem. Powerfulness is used to locate a constrained optimum within these sublevels and to obtain the common direction; it is not used as a replacement for either curvature or sampling access.

If $p_{\theta_0}(S) \le \alpha$, θ_0 is already a constrained optimum. The nontrivial case has $p_{\theta_0}(S) > \alpha$. Approximate initialization and numerical evaluation can be used if their errors are included in the tolerances below.

4.1 The common direction and the boundary proxy

Proposition 9 (Common non-increasing direction)**.** *Under powerfulness, common support, and the global covariance upper bound of Theorem 8, every bounded $B \subseteq \Theta$ admits a unit vector u such that $u^\top \nabla F_S(\theta) \le 0$ for all $\theta \in B$. Neither closedness of B nor a minimum of F_S over B is required. For a uniformly random unit vector v, $u^\top v \ge a_0/\sqrt{k}$ with an absolute positive probability, for a suitable absolute $a_0 > 0$.*

Fix any ϑ with positive invalid mass. Powerfulness supplies θ_j with $p_{\theta_j}(S) \downarrow 0$ and $F_S(\theta_j) \le F_S(\vartheta)$. The global covariance upper bound prevents the mass of S from approaching zero along a bounded parameter sequence. Hence $\|\theta_j\| \to \infty$. Pass to a convergent subsequence of directions and use convexity:

$$\left\langle \nabla F_S(\theta), \frac{\theta_j - \theta}{\|\theta_j - \theta\|_2} \right\rangle \le \frac{F_S(\vartheta) - F_S(\theta)}{\|\theta_j - \theta\|_2} \longrightarrow 0.$$

The normalized directions have the same limit for all θ in a bounded set. Appendix C supplies the mass bound and the random-direction argument.

Proposition 10 (Approximate minimizers near the boundary)**.** *Under the curvature assumptions of Theorem 8, let C_c be compact, $c > 0$, and let u, v be unit vectors with $u^\top \nabla F_S \le 0$ on C_c and $u^\top v \ge \gamma > 0$. Fix $\rho > 0$ and $\xi \ge 0$. If $\widehat{\theta} \in C_c$ satisfies $F_{\rho,v}(\widehat{\theta}) - \min_{C_c} F_{\rho,v} \le \xi$, then*

$$0 \le \mathcal{L}_0 + c - \mathcal{L}(\widehat{\theta}) \le \frac{L\sqrt{2c/\lambda}}{\rho\gamma}\,\xi, \qquad F_S(\widehat{\theta}) - m(c) \le \xi + \rho\,\mathrm{diam}(C_c). \tag{11}$$

Indeed, move from $\widehat{\theta}$ along u until hitting the boundary. Along this ray the proxy decreases at rate at least $\rho\gamma$, so its length is at most $\xi/(\rho\gamma)$. On C_c, $\|\nabla\mathcal{L}\| \le L\sqrt{2c/\lambda}$. The second bound follows by

Algorithm 1 PSGD for a boundary proxy

Require: Sublevel c, unit direction v, perturbation ρ, iterations N, valid-data mean $\widehat{\mu}$.
1: $w_1 \leftarrow \theta_0$; $h \leftarrow D/(G_\rho\sqrt{N})$.
2: **for** $t = 1, \dots, N$ **do**
3: Draw $Z_t \sim p^S_{w_t}$ using Sample_S.
4: $g_t \leftarrow T(Z_t) - \widehat{\mu} - \rho v$.
5: $w_{t+1} \leftarrow \Pi_{C_c}(w_t - h g_t)$.
6: **end for**
7: **return** $N^{-1}\sum_{t=1}^{N} w_t$.

bounding the change in the linear perturbation over C_c. In particular an exact minimizer of F_S can be moved to the boundary without increasing F_S.

Projected stochastic gradient descent (PSGD) uses $T(Z) - \widehat{\mu} - \rho v$, with $Z \sim p^S_\theta$, as an unbiased gradient of the proxy. Averaging the iterates gives the convex rate $O(DG_\rho/\sqrt{N})$, where D bounds the diameter and G^2_ρ bounds the proxy-gradient second moment. Repeated random directions, a boundary-slack check, and objective comparisons produce a candidate satisfying both $F_S(\widehat{\theta}) - m(c) \le \eta$ and $\mathcal{L}_0 + c - \mathcal{L}(\widehat{\theta}) \le \zeta$. Appendix D gives the error budgets and oracle counts.

The parameters can be chosen explicitly. Write $\ell = 1 + \log(1/\alpha)$. On the search region let $B_L = L\sqrt{2H/\lambda}$ and let $a_0/\sqrt{k}$ be the required alignment. For objective tolerance η and boundary tolerance ζ, take

$$\rho = \frac{\eta}{8D}, \qquad \xi = \min\left\{\frac{\eta}{8}, \frac{\rho a_0 \zeta}{2\sqrt{k} B_L}\right\}, \qquad N \ge C\frac{D^2 G^2_\rho}{\xi^2},$$

where G^2_ρ bounds the second moment of the proxy gradient. The truncated estimates give such a bound polynomial in k, K, L, H, λ^{-1} and ℓ, as proved in Lemma 15. The first choice keeps the perturbation error below the objective tolerance; the second controls the distance to the boundary; the third makes the PSGD error sufficiently small with constant probability.

Each run uses a fresh random direction. Its alignment need not be observed: the algorithm checks the NLL boundary slack of the output and compares objective values among the outputs that pass this check. A suitably aligned and accurate run passes and has a small objective gap. An output selected over that run has a similarly small gap, even if its own direction was unsuitable. Repeating $O(\log(Q/\delta))$ times makes this argument hold for every one of the Q sublevel solves. Appendix K proves the repetition and selection bounds and quantifies numerical projection errors.

4.2 The KL bound and the outer search

Lemma 11 (Conditional KL across neighboring sublevels)**.** *Let a minimize F_S on a smaller sublevel and let b minimize it on a larger sublevel C containing a and $\widehat{\theta}$. Suppose $\mathrm{diam}(C) \le D$, F_S is M_S-smooth on C, $F_S(\widehat{\theta}) - F_S(b) \le \eta$, and $\Delta = F_S(a) - F_S(b) \ge 0$. Then*

$$\mathrm{KL}(p^S_a \| p^S_{\widehat{\theta}}) \le \eta + D\sqrt{2M_S\Delta}. \tag{12}$$

If $\widehat{\theta}$ belongs to the same sublevel minimized by a, first-order optimality gives the simpler bound $\mathrm{KL}(p^S_a \| p^S_{\widehat{\theta}}) \le F_S(\widehat{\theta}) - F_S(a)$.

The exact identity includes the linear term

$$\mathrm{KL}(p^S_a \| p^S_{\widehat{\theta}}) = F_S(\widehat{\theta}) - F_S(a) - \langle \nabla F_S(a), \widehat{\theta} - a\rangle. \tag{13}$$

The proof in Appendix E bounds its negative by $\Delta + D\sqrt{2M_S\Delta}$ using a feasible segment. Thus, when the returned point is in a larger sublevel, the objective gap must be controlled to order ε^2.

On the search region one can take

$$g = \min\{1/4, c_1\lambda/(K^2\ell^2)\}, \quad D^2 = 8H/\lambda, \quad M_S = CK^2\ell^2, \tag{14}$$

Algorithm 2 Nested search with mass checks

Require: Fixed family, valid data, $\alpha, \varepsilon, \delta$; scales in (14)–(15).
Form $\widehat{\mu}$ and find θ_0.
Apply the initial threshold test of Lemma 18; return θ_0 if the test accepts.
Set the certified left threshold to 0 with candidate θ_0.
for $c = g, 2g, \ldots, \lceil \ell/g \rceil g$ **do**
  Obtain $w_c \in C_c$ with objective error η and boundary slack ζ.
  Compute $\widehat{z}_c$ with $|\widehat{z}_c - \log p_{w_c}(S)| \le \nu$; put $h = \eta + \zeta$.
  if $\widehat{z}_c - h - \nu > \log \alpha$ **then**
    Set the certified left threshold to c and continue.
  else if $\widehat{z}_c + \nu < \log \alpha$ **then**
    Set the certified right threshold to c and **break**.
  else
    return w_c (its log mass is within $h + 2\nu$ of $\log \alpha$).
  end if
end for
Bisect the certified bracket, applying the same three-way test at each midpoint.
Stop also when bracket width and a certified upper bound on its exact m-gap are both at most q; return the right candidate.

with a sufficiently small absolute c_1. Choose

$$q = \min\left\{\frac{\varepsilon}{64}, \frac{\varepsilon^2}{4096 \max\{1, M_S D^2\}}, \frac{g}{64}\right\}, \qquad \eta = \zeta = \nu = q/16. \tag{15}$$

Here η is objective accuracy, ζ is boundary slack, and ν is log-mass estimation error. When $K^2 = O(L)$, the accuracy scale is $O(g\varepsilon^2/\ell)$ in the nonconstant regime. The runtime accounts for this tighter tolerance and the convex PSGD rate.

The bracket refers to the monotone exact quantity $\log r(c)$, not noisy output masses. Small successive gaps give the progress bound of Proposition 7; they do not bypass the mass test. At a crossing, Lemma 11 applies with the exact objective gap between the target and the larger sublevel. For the final bracket, that gap is at most q. Appendix F proves termination, the $\Omega(\alpha)$ mass bound throughout the search, and all three guarantees in (10).

4.3 How the proof gives all three guarantees

The boundary construction links the convex objective to the original constraint. We describe that link and explain why the outer search needs both objective and mass estimates. The full bounds, including initialization and stopping, are proved in Appendix F.

Identifying the target sublevel. Assume $p_{\theta_0}(S) > \alpha$, since otherwise θ_0 is optimal. Powerfulness provides a feasible parameter with NLL at most $\mathcal{L}_0 + \log[p_{\theta_0}(S)/\alpha]$. Compactness then gives a constrained optimum a within C_H. Its invalid mass must equal α: if it were smaller, moving a short distance toward θ_0 would preserve feasibility and strictly improve NLL. Set $c_* = \mathcal{L}(a) - \mathcal{L}_0$. The same segment argument shows that every point of C_{c_*} has invalid mass at least α.

Moreover, a minimizes F_S on C_{c_*}. To see this, suppose $F_S(b) < F_S(a)$ for some $b \in C_{c_*}$. If b is infeasible, powerfulness gives a feasible b' with

$$\mathcal{L}(b') \le \mathcal{L}(b) + \log \frac{p_b(S)}{\alpha} = F_S(b) - \log \alpha < \mathcal{L}(a),$$

contradicting optimality. A feasible b gives the same contradiction directly or by the preceding segment argument. This identifies the target in the KL guarantee without assuming that the observed data come from a model conditioned on S.

Why the boundary is needed. At an exact boundary minimizer b_c, the two quantities $m(c)$ and $\mathcal{L}_0 + c$ determine the mass: $\log p_{b_c}(S) = m(c) - \mathcal{L}_0 - c$. For a candidate w_c with objective error e_c and boundary slack s_c, the exact relation is

$$\log p_{w_c}(S) = \log r(c) + e_c + s_c, \quad e_c = F_S(w_c) - m(c), \quad s_c = \mathcal{L}_0 + c - \mathcal{L}(w_c).$$

Controlling only e_c leaves s_c uncontrolled. The linear proxy supplies the missing bound: motion in the common direction decreases the proxy at rate at least $\rho\gamma$, so a point with proxy error ξ is at distance at most $\xi/(\rho\gamma)$ from the boundary along that direction. The NLL gradient bound converts this distance to the required bound on s_c. The perturbation changes objective comparisons by at most ρD, which is included in η.

Keeping the truncated oracle applicable. The grid moves to the next sublevel only after establishing $r(c) > \alpha$. Its first crossing of the target therefore lies within g of c_*. Strong convexity places every point in C_{c_*+g} within distance $\sqrt{2g/\lambda}$ of C_{c_*}. The truncated mean estimate controls how quickly log mass changes along the intervening segment:

$$\|\nabla \log p_\theta(S)\|_2 \le CK[1 + \log(1/p_\theta(S))].$$

Integrating this inequality gives the mass lower bound α/e for the choice of g in (14). Thus every inner iterate remains within the domain of the $\Omega(\alpha)$-truncated oracle. Bisection then stays inside the same region.

Converting the stopping bounds to KL. When the returned point lies in C_{c_*}, optimality at a makes the linear term in (13) nonnegative, so the objective gap bounds KL directly. A point in a larger sublevel needs the additional argument in Lemma 11. If b minimizes F_S there, set $\Delta = F_S(a) - F_S(b)$. Smoothness along the feasible segment from a to the returned point bounds the negative linear term by $\Delta + D\sqrt{2M_S\Delta}$. The Δ term cancels against the corresponding objective decrease, leaving $\eta + D\sqrt{2M_S\Delta}$. This explains the quadratic dependence on ε in the objective tolerance.

The stopping rule controls Δ either through a mass estimate close to α or through the final bracket's objective gap. In the first case, writing $\kappa = \eta + \zeta + 2\nu$, the proof gives an NLL excess at most $\eta + \kappa$, invalid mass at most αe^κ, and $\Delta \le \eta + \zeta + \kappa$. In the second case, the bracket bounds give an NLL excess at most q and $\Delta \le q$. The choices in (15) make each error at most its stated tolerance. The appendix also proves that bisection reaches these conditions in polynomially many calls, rather than assuming a stopping rule will eventually succeed.

Accounting for the two sampling costs. A conditional sample supplies the stochastic gradient, whereas an ordinary sample together with an invalidity query supplies a Bernoulli observation of the mass. These tasks have different costs. For mass bounded below by α/e, estimating its logarithm to error ν with failure probability δ' requires

$$N_{\text{mass}} = O\left(\frac{1}{\alpha\nu^2}\log\frac{1}{\delta'}\right)$$

ordinary samples. This also estimates F_S after evaluating $\mathcal{L}$. A sublevel solve with R repetitions and N PSGD iterations uses RN conditional calls. If these are implemented by rejection sampling, their expected ordinary-sampling cost is at most eRN/α. The geometric waiting-time bound in Appendix K gives a corresponding high-probability bound. Thus the logarithmic dependence on $1/\alpha$ for conditional-oracle calls does not imply logarithmic dependence for all invalidity queries. Both costs, including the smaller objective tolerance needed for KL, are included in the theorem.

Scope. The result assumes a noiseless validity oracle. It concerns optimization with that access, rather than verification of factual truth without it. No unimodality assumption is imposed on the output distribution: convexity is in natural parameters. The dimension and conditioning constants can still make high-dimensional instances costly. Powerfulness, tail bounds, conditioning, and computational access are separate assumptions on the initial family. The Gaussian illustration in Appendix H shows the likelihood–validity tradeoff; it does not establish these assumptions for every Gaussian model.

AI USE STATEMENT

Generative AI was used to assist with manuscript revision, literature lookup, reference formatting, notation and language checks. The authors are responsible for the final content, including AI-assisted text and mathematical claims.

REFERENCES

D. Angluin. Learning Regular Sets from Queries and Counterexamples. *Information and Computation*, 75(2):87–106, 1987.

Andrew R. Barron and Chyong-Hwa Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991. doi: 10.1214/aos/1176348252. URL https://projecteuclid.org/euclid.aos/1176348252.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003. URL http://www.jmlr.org/papers/v3/bengio03a.html.

Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 639–649, 2018. URL https://api.semanticscholar.org/CorpusID:52189557.

Constantinos Daskalakis, Dhruv Rohatgi, and Emmanouil Zampetakis. Truncated linear regression in high dimensions. *Advances in Neural Information Processing Systems*, 33:10338–10347, 2020.

Steve Hanneke, Adam Tauman Kalai, Gautam Kamath, and Christos Tzamos. Actively avoiding nonsense in generative models. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 209–227. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/hanneke18a.html.

J. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440, 1997.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pp. 160–171, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703836. doi: 10.1145/3618260.3649777. URL https://doi.org/10.1145/3618260.3649777.

Konstantinos Karatapanis, Vasilis Kontonis, and Christos Tzamos. Oracle efficient truncated statistics. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1983888547340c64635a29c45cef04e8-Abstract-Conference.html.

Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Jane Lee, Andre Wibisono, and Manolis Zampetakis. Learning exponential families from truncated samples. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=PxcWJqO3qj.

Nick Rittler and Kamalika Chaudhuri. Distribution learning with valid outputs beyond the worst-case. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 22850–22877. Curran Associates, Inc., 2024. doi: 10.52202/079017-0719. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2897c21877ff242e9831ec112c95f6bb-Paper-Conference.pdf.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(183):1–59, 2017. URL https://jmlr.org/papers/v18/16-011.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

A A CONSTRUCTION ESTABLISHING THE CAPACITY REQUIREMENT

Proposition 12 (Selective augmentation)**.** *Let $p_\theta(x) = h(x)e^{\theta^\top T(x) - A(\theta)}$ on X, and let $x^* \notin X$ be a designated valid output. On $X' = X \cup \{x^*\}$ use the original reference measure together with a unit atom at x^*. Fix $h_* > 0$. Extend $T(x^*) = 0$ and use the sufficient statistics $(T, \mathbf{1}_S, \mathbf{1}_{\{x^*\}})$. The resulting family has unnormalized weights*

$$\widetilde{q}_{\theta,a,b}(x) = \begin{cases} h(x)e^{\theta^\top T(x) + a\mathbf{1}_S(x)}, & x \in X, \\ h_* e^b, & x = x^*. \end{cases} \tag{16}$$

This fixed augmented family is powerful uniformly over valid data in $V \cup \{x^*\}$. It contains the original family in its closure as $a = 0$ and $b \to -\infty$.

Proof. First take an original member with $s = p_\theta(S) > \alpha$. Choose

$$a = \log(\alpha/s), \qquad h_* e^b = (s - \alpha)e^{A(\theta)}.$$

The normalizer in (16) remains $e^{A(\theta)}$, since the three parts have weights $e^{A(\theta)}(1 - s)$, $e^{A(\theta)}\alpha$, and $e^{A(\theta)}(s - \alpha)$. Consequently the normalized density is

$$q(x) = \begin{cases} p_\theta(x), & x \in V, \\ (\alpha/s)p_\theta(x), & x \in S, \\ s - \alpha, & x = x^*. \end{cases} \tag{17}$$

This preserves the density at every original valid point, sets invalid mass to α, and preserves the conditional distribution on S.

It remains to check powerfulness for every finite-parameter member of the augmented family, rather than just for original members. Let $q = q_{\theta,a,b}$ have normalizer Z and invalid mass $s > \alpha$. Keep θ fixed and set

$$a' = a + \log(\alpha/s), \qquad h_* e^{b'} = h_* e^b + (s - \alpha)Z.$$

Removing $(s - \alpha)Z$ from the invalid part and adding it to the atom leaves the normalizer equal to Z. Densities on V stay fixed and the valid atom's mass increases. Hence every valid observation has nondecreasing density, while $q'(S)/q(S) = \alpha/s < 1$. This proves (2) for every valid data set. Finally, taking $a = 0$, $b \to -\infty$ recovers each original member in total variation. The zero mass at the new atom is a limit, not a finite natural parameter. □

This construction explains how added expressivity can establish the capacity condition when selecting a model. It does not claim that the original and augmented families have the same covariance, sampling complexity, or parameter geometry. In particular, the variance of the added indicator is $q(S)(1 - q(S))$, which tends to zero as $q(S) \to 0$. Thus a global positive covariance lower bound does not follow from the construction. The optimization theorem starts with a family for which powerfulness and the additional assumptions have already been established.

B PROOFS OF THE TRUNCATED ESTIMATES

Proof of Lemma 5. Fix a unit u and write $Y = u^\top(T - \mathbb{E}_{p_\theta} T)$. Conditional on S,

$$\Pr(|Y| > t \mid S) \le \min\{1, 2e^{-t/K}/s\}.$$

Set $t_0 = K \log(2/s)$. Integrating this bound gives

$$\mathbb{E}[|Y| \mid S] \le t_0 + K,$$

$$\mathbb{E}[Y^2 \mid S] \le t_0^2 + 2K(t_0 + K).$$

Taking the supremum over unit u proves the mean bound. The conditional variance is no greater than the second moment about the original mean, proving the covariance bound. For the higher moments, let E be an exponential random variable of mean K. The conditional tail bound shows that $|Y|$ under S is stochastically dominated by $t_0 + E$. For every integer $q \ge 1$, Minkowski's inequality gives

$$(\mathbb{E}[|Y|^q \mid S])^{1/q} \le t_0 + K(q!)^{1/q} \le K[\log(2/s) + q].$$

For $W = Y - \mathbb{E}[Y \mid S]$, centering increases this bound by at most a factor two. Since $\ell_s \ge 1$ and $q^q \le e^q q!$, enlarging an absolute constant gives

$$\mathbb{E}[|W|^q \mid S] \le q! B^q, \qquad B = CK\ell_s, \qquad q \ge 2. \tag{18}$$

For $|z|B \le 1/2$, expand the exponential and use $\mathbb{E}[W \mid S] = 0$:

$$\mathbb{E}[e^{zW} \mid S] \le 1 + \sum_{q \ge 2} (|z|B)^q \le 1 + 2z^2B^2 \le e^{2z^2B^2}.$$

The series is justified by the exponential tail and absolute convergence. For independent copies $W_1, \dots, W_N$, the exponential Markov inequality therefore gives, for $0 < z \le 1/(2B)$,

$$\Pr\left(\frac{1}{N}\sum_i W_i > v\right) \le \exp[-N(zv - 2z^2B^2)].$$

Taking $z = \min\{v/(8B^2), 1/(4B)\}$ and applying the same argument to $-W$ proves

$$\Pr\left(\left|\frac{1}{N}\sum_i W_i\right| > v\right) \le 2\exp\left[-cN\min\left\{\frac{v^2}{B^2}, \frac{v}{B}\right\}\right]. \tag{19}$$

With $N = 1$, this implies a sub-exponential tail of scale CB; for small v/B one uses the trivial probability bound one. Apply (19) to each coordinate of the centered statistic with $v = t/\sqrt{k}$. If all coordinate errors are at most this value, the Euclidean error is at most t. A union bound over k coordinates yields the sample bound in the lemma. This proves the moment, covariance, tail, and estimation claims under the stated uniform scale, without an additional concentration assumption. □

A covariance bound at one parameter does not by itself imply concentration. A sufficient condition is a uniform covariance upper bound in a neighborhood of fixed radius in the natural parameter space. Indeed, if $\theta + tu \in \Theta$ for $|t| \le r$ and $\nabla^2 A \preceq LI$ along these tilts, then

$$\log \mathbb{E}_{p_\theta} e^{tu^\top (T - \mathbb{E}_{p_\theta} T)} = A(\theta + tu) - A(\theta) - tu^\top \nabla A(\theta) \le Lt^2/2.$$

Chernoff's inequality gives a uniform sub-exponential scale $K = O(\sqrt{L} + r^{-1})$. The radius must be uniform over the region in question. We do not infer it merely from openness of Θ at each individual parameter.

C Powerfulness and the geometry of the target

Proof of Proposition 9. Fix $\vartheta \in \Theta$ with $p_\vartheta(S) > 0$. For $\alpha_j \downarrow 0$ below this mass, powerfulness supplies θ_j satisfying $p_{\theta_j}(S) = \alpha_j$ and

$$\mathcal{L}(\theta_j) - \mathcal{L}(\vartheta) \le \log \frac{p_\vartheta(S)}{\alpha_j}, \qquad F_S(\theta_j) \le F_S(\vartheta).$$

We verify that this sequence escapes every bounded parameter set. Choose $B_0 < \infty$ such that $s_0 = p_\vartheta(S \cap \{\|T\|_2 \le B_0\}) > 0$; this is possible by positive mass and almost-everywhere finiteness of T. If $\|\theta - \vartheta\|_2 \le R$, the global Hessian upper bound and convexity of Θ imply

$$A(\theta) - A(\vartheta) \le R\|\nabla A(\vartheta)\|_2 + LR^2/2.$$

The likelihood ratio on this subset of S therefore gives

$$p_\theta(S) \ge s_0 \exp[-RB_0 - R\|\nabla A(\vartheta)\|_2 - LR^2/2] > 0. \tag{20}$$

This bound is uniform on each bounded parameter ball, even if that ball meets the boundary of Θ. Thus $\|\theta_j - \vartheta\|_2 \to \infty$.

By compactness of the unit sphere, pass to a subsequence on which $(\theta_j - \vartheta)/\|\theta_j - \vartheta\|_2 \to u$. For bounded B, $(\theta_j - \theta)/\|\theta_j - \theta\|_2 \to u$ uniformly in $\theta \in B$. Convexity gives, for each such θ,

$$\left\langle \nabla F_S(\theta), \frac{\theta_j - \theta}{\|\theta_j - \theta\|_2} \right\rangle \le \frac{F_S(\theta_j) - F_S(\theta)}{\|\theta_j - \theta\|_2} \le \frac{F_S(\vartheta) - F_S(\theta)}{\|\theta_j - \theta\|_2} \longrightarrow 0.$$

No minimizer on B is used. For the probabilistic statement, let $G \sim \mathcal{N}(0, I_k)$ and $v = G/\|G\|_2$. The event $u^\top G \ge 1/2$ has absolute positive probability, and Markov's inequality gives $\Pr(\|G\|_2 > 3\sqrt{k}) \le 1/9$. Their intersection has absolute positive probability, and there $u^\top v \ge 1/(6\sqrt{k})$. □

Lemma 13 (Target sublevel)**.** *Suppose the assumptions of Theorem 8 hold and $p_{\theta_0}(S) > \alpha$. Then a constrained optimum $a = \theta^\star_\alpha$ exists, has $p_a(S) = \alpha$, and satisfies*

$$c_* = \mathcal{L}(a) - \mathcal{L}_0 \le \log \frac{p_{\theta_0}(S)}{\alpha} < \ell.$$

It minimizes F_S on C_{c_*}. For every $0 \le c \le H$, there is a boundary minimizer b_c; $r(c_*) = \alpha$ and $p_\theta(S) \ge \alpha$ throughout C_{c_*}.

Proof. Powerfulness at θ_0 gives a feasible point with NLL at most $\mathcal{L}_0 + \log(p_{\theta_0}(S)/\alpha)$. The feasible portion of this compact sublevel is closed and nonempty, so the optimum is attained there. If its mass were strictly below α, continuity along the segment toward the unique NLL minimizer θ_0 would give a feasible point of smaller NLL. This proves the active constraint. The same argument shows that no point of C_{c_*} has mass below α.

Suppose $b \in C_{c_*}$ has $F_S(b) < F_S(a)$. If $p_b(S) \le \alpha$, it contradicts either optimality directly or the preceding segment argument. Otherwise powerfulness produces a feasible b' with

$$\mathcal{L}(b') \le \mathcal{L}(b) + \log[p_b(S)/\alpha] = F_S(b) - \log\alpha < \mathcal{L}(a),$$

again a contradiction. Thus a minimizes F_S on C_{c_*}.

On any compact C_c with $c > 0$, take a minimizer of F_S and the direction from Proposition 9. If the minimizer is interior, move along this direction until reaching the boundary. Compact containment in Θ ensures the ray exits through $\mathcal{L} = \mathcal{L}_0 + c$; F_S does not increase along the segment. For $c = 0$, $C_0 = \{\theta_0\}$. This proves the final claims using (8). □

Lemma 14 (Continuity of nested minima and mass on a larger sublevel)**.** *On C_H, $\operatorname{diam}(C_H) \le D = 2\sqrt{2H/\lambda}$. If $\|\nabla F_S\| \le B$ on C_d, then for $0 \le c < d \le H$,*

$$0 \le m(c) - m(d) \le B\sqrt{2(d-c)/\lambda}. \tag{21}$$

For g as in (14), every parameter of C_{c_*+g} assigns at least α/e mass to S.

Proof. Strong convexity and $\nabla\mathcal{L}(\theta_0) = 0$ give the diameter bound. Let $x \in C_d \setminus C_c$ and let y be the intersection of $[\theta_0, x]$ with the boundary of C_c. The radial derivative of $\mathcal{L}$ at y is nonnegative. Strong convexity therefore gives

$$\tfrac{\lambda}{2}\|x - y\|_2^2 \le \mathcal{L}(x) - \mathcal{L}(y) \le d - c.$$

Apply this to a minimizer x of F_S on C_d and use the gradient bound on $[y, x]$ to obtain (21).

For the mass bound, project radially from $x \in C_{c_*+g}$ to $y \in C_{c_*}$, unless x already belongs to the latter set. The distance is at most $\sqrt{2g/\lambda}$ and $p_y(S) \ge \alpha$. By Lemma 5,

$$\|\nabla \log p_\theta(S)\|_2 = \|\mathbb{E}_{p^S_\theta} T - \mathbb{E}_{p_\theta} T\|_2 \le CK[1 + \log(1/p_\theta(S))].$$

Along the segment, this differential inequality implies

$$1 + \log(1/p_x(S)) \le \ell \exp(CK\sqrt{2g/\lambda}).$$

For sufficiently small c_1 in (14), the right-hand side is at most $\ell + 1$. Hence $p_x(S) \ge \alpha/e$. This argument uses the uniform tail assumption on the entire segment, rather than assuming the desired mass bound in advance. □

D Optimization steps and their costs

Lemma 15 (Stochastic gradient second moment)**.** *On a sublevel $C_c \subseteq C_{c_*+g}$, $Z \sim p_\theta^S$ gives the unbiased gradient $G_\theta = T(Z) - \widehat{\mu}$. Uniformly on this region,*

$$\mathbb{E}\|G_\theta\|_2^2 = \operatorname{tr}\operatorname{Cov}_{p_\theta^S}(T) + \|\mathbb{E}_{p_\theta^S} T - \widehat{\mu}\|_2^2 \le C(k+1)K^2\ell^2 + 4L^2H/\lambda =: G^2. \tag{22}$$

For the proxy gradient $G_\theta - \rho v$ the bound $G_\rho^2 = 2G^2 + 2\rho^2$ suffices.

Proof. Unbiasedness follows from $\nabla F_S = \mathbb{E}_{p_\theta^S} T - \widehat{\mu}$. Decompose the mean difference into $\mathbb{E}_{p_\theta^S} T - \mathbb{E}_{p_\theta} T$ and $\nabla\mathcal{L}(\theta)$. The first is at most $CK\ell$ and the second at most $L\|\theta - \theta_0\| \le L\sqrt{2H/\lambda}$. Apply Lemma 5 and $(a+b)^2 \le 2a^2 + 2b^2$. □

Theorem 16 (Convex PSGD)**.** *Let f be convex and differentiable on a compact convex set C of diameter at most $D > 0$, and let $w_1 \in C$. Suppose the stochastic gradients are conditionally unbiased with conditional second moment at most G_f^2, where $G_f > 0$. For $N \ge 1$, with step size $h = D/(G_f\sqrt{N})$ and projected updates $w_{t+1} = \Pi_C(w_t - hg_t)$, the average $\overline{w} = N^{-1}\sum_{t=1}^N w_t$ satisfies*

$$\mathbb{E}[f(\overline{w}) - \min_C f] \le DG_f/\sqrt{N}.$$

Proof. For a minimizer w^*, projection and expansion of the squared norm give

$$2h\,\mathbb{E}[f(w_t) - f(w^*)] \le \mathbb{E}[\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2] + h^2G_f^2.$$

Sum over t, divide by $2hN$, and use convexity at the average to obtain $D^2/(2hN) + hG_f^2/2$. Substitution proves the claim. The objective f need not be strongly convex. □

Producing a boundary candidate. Given objective tolerance η and boundary tolerance ζ, let $a_0/\sqrt{k}$ be the alignment bound in Proposition 9, and put $B_L = L\sqrt{2H/\lambda}$. Take

$$\rho = \frac{\eta}{8D}, \qquad \xi = \min\left\{\frac{\eta}{8}, \frac{\rho a_0 \zeta}{2\sqrt{k}B_L}\right\}, \qquad N \ge CD^2G_\rho^2/\xi^2.$$

By Theorem 16 and Markov's inequality, a run has proxy error at most ξ with arbitrarily high fixed constant probability, by choosing the absolute C large enough. Independently draw v each time. A run with both this accuracy and the required alignment has objective error at most $\eta/4$ and boundary slack at most $\zeta/2$. Their joint probability is bounded below by an absolute positive constant. Repeat $R = O(\log(Q/\delta))$ times, where Q bounds all sublevel calls, discard candidates whose measured boundary slack exceeds ζ, and choose the remaining candidate with the smallest estimated F_S. With objective evaluation error at most $\eta/8$, the selected candidate has objective error below η. A candidate with inaccurate PSGD cannot invalidate this selection bound: the selected value is compared against an accurate surviving candidate. If no candidate survives, return θ_0 and declare failure; this event is included in the failure probability. Numerical NLL, projection, and boundary errors use the remaining halves of these budgets.

Estimating masses and objective values. If $p_\theta(S) \ge c_0\alpha$, independent ordinary samples give Bernoulli invalidity observations. Multiplicative Chernoff bounds imply that

$$N_{\text{mass}} = O\left(\frac{\log(1/\delta')}{\alpha\nu^2}\right) \tag{23}$$

samples suffice for $|\widehat{z} - \log p_\theta(S)| \le \nu$, for $0 < \nu < 1/2$. The estimate of F_S is $\mathcal{L} + \widehat{z}$. These are ordinary sampling and membership costs, additional to the truncated calls in PSGD. If using rejection sampling for PSGD as well, every conditional draw costs at most e/α proposals in expectation on the search region. Standard geometric tail bounds supply a polynomial high-probability budget.

E Conditional KL and the linear term

Proof of Lemma 11. The exponential-family likelihood ratio on S gives

$$\mathrm{KL}(p_a^S \| p_{\widehat{\theta}}^S) = A_S(\widehat{\theta}) - A_S(a) - \nabla A_S(a)^\top(\widehat{\theta} - a).$$

The affine empirical-data terms cancel, proving (13). When $\widehat{\theta}$ lies in the sublevel minimized by a, first-order optimality makes the inner product nonnegative.

For the larger sublevel, set $d = \widehat{\theta} - a$ and $\Delta = F_S(a) - F_S(b)$. For $0 < t \le 1$, $a + td \in C$, and smoothness together with optimality at b gives

$$F_S(b) \le F_S(a + td) \le F_S(a) + t\nabla F_S(a)^\top d + \frac{M_S t^2 D^2}{2}.$$

Thus

$$-\nabla F_S(a)^\top d \le \frac{\Delta}{t} + \frac{M_S D^2 t}{2} \le \Delta + D\sqrt{2M_S\Delta},$$

where the last inequality follows by taking $t = \min\{1, \sqrt{2\Delta/(M_S D^2)}\}$, with the zero cases obtained by a limit. Since $F_S(\widehat{\theta}) - F_S(a) \le \eta - \Delta$, substitution into the exact identity proves (12). □

Suppose successive approximate sublevel minima differ in objective by at most τ, each has error at most η, and the target threshold is between those sublevels. Then the exact neighboring gap is at most $\tau + 2\eta$; objective estimation errors must be added if the comparison is estimated. Hence (12) gives

$$\mathrm{KL}(p_a^S \| p_{\widehat{\theta}}^S) \le \eta + D\sqrt{2M_S(\tau + 2\eta)}.$$

For $D^2 = O(\ell/\lambda)$, $M_S = O(L(1+\ell^2))$, and $g = O(\lambda/(L\ell^2))$ in the regime $K^2 = O(L)$, choosing sufficiently small $\eta, \tau = O(g\varepsilon^2/\ell)$ yields $O(\varepsilon)$ KL error. This bounds conditional KL when the target threshold lies between the two sublevels. The invalid mass must also be checked before returning a solution.

F Outer search and proof of the main theorem

We give the bounds maintained during the search, its stopping criteria, and the required accuracy for Algorithm 2. Let Q bound the number of sublevel solves, as specified below. Allocate failure probability $\delta/[2(Q+1)]$ to initialization and each solve. Within each solve, divide this budget between finding a successful candidate and estimating objective values; individual unsuccessful PSGD runs are allowed. If rejection sampling is used, reserve the remaining $\delta/2$ for its proposal limits. A union bound gives probability at least $1 - \delta$ that every selected candidate and required estimate has the stated accuracy. The argument below holds on this event.

Initialization. Let $\widehat{p}_0$ estimate $p_{\theta_0}(S)$ from $O(\alpha^{-1}\nu^{-2}\log(Q/\delta))$ Bernoulli samples. Return θ_0 if $\widehat{p}_0 \le \alpha e^{2\nu}$. Otherwise Lemma 18 certifies $p_{\theta_0}(S) > \alpha$ and log-estimation error at most ν. With a constant-factor adjustment to the sample count, multiplicative and threshold Chernoff bounds imply that a returned parameter either has $p_{\theta_0}(S) \le \alpha$ or has $\log[p_{\theta_0}(S)/\alpha] \le 4\nu$. In the former case θ_0 is itself a constrained optimum. In the latter case Lemma 13 applies, $\theta_0 \in C_{c_*}$, and

$$F_S(\theta_0) - F_S(a) = -c_* + \log[p_{\theta_0}(S)/\alpha] \le 4\nu.$$

First-order optimality on C_{c_*} proves the KL guarantee, NLL is no larger than the optimum, and the mass bound follows from $4\nu \ll \varepsilon$. This threshold test does not attempt a relative estimate of an arbitrarily tiny mass.

Checking the candidate mass. For any successful boundary candidate w_c, write

$$e_c = F_S(w_c) - m(c) \in [0, \eta], \qquad s_c = \mathcal{L}_0 + c - \mathcal{L}(w_c) \in [0, \zeta].$$

Then

$$\log p_{w_c}(S) = \log r(c) + e_c + s_c. \tag{24}$$

Writing $h = \eta + \zeta$ and allowing log-estimation error ν, the test $\widehat{z}_c - h - \nu > \log \alpha$ certifies $r(c) > \alpha$, hence $c < c_*$. The test $\widehat{z}_c + \nu < \log \alpha$ certifies $r(c) < \alpha$, hence $c > c_*$. If neither holds, the candidate satisfies

$$|\log[p_{w_c}(S)/\alpha]| \le h + 2\nu =: \kappa. \tag{25}$$

The grid advances only when the test establishes $c < c_*$. Its first crossing or ambiguous point is therefore at most $c_* + g$. Since $c_* < \ell$ and $\log r(c)$ decreases by at least the threshold increment, a crossing or ambiguous point occurs by the final grid threshold. Lemma 14 ensures that every parameter accessed, including every PSGD iterate, has mass at least α/e. All subsequent bisection thresholds are smaller than the first right threshold, preserving this invariant and the $\Omega(\alpha)$ oracle convention.

Returning a candidate with mass close to α. Suppose (25) holds at threshold c. If $c \le c_*$, $w_c \in C_{c_*}$ and

$$F_S(w_c) - F_S(a) = \mathcal{L}(w_c) - \mathcal{L}(a) + \log[p_{w_c}(S)/\alpha] \le \kappa.$$

The NLL guarantee is immediate, and optimality on C_{c_*} gives KL divergence at most κ. If $c > c_*$, let $b = b_c$ and $\Delta = F_S(a) - F_S(b) \ge 0$. Since $F_S(w_c) \le F_S(a) + \eta$,

$$\mathcal{L}(w_c) - \mathcal{L}(a) \le \eta + \kappa.$$

Also $\mathcal{L}(w_c) \ge \mathcal{L}_0 + c - \zeta \ge \mathcal{L}(a) - \zeta$, so $F_S(w_c) - F_S(a) \ge -\zeta - \kappa$. Consequently

$$\Delta \le \eta + \zeta + \kappa, \qquad \mathrm{KL}(p_a^S \| p_{w_c}^S) \le \eta + D\sqrt{2M_S(\eta + \zeta + \kappa)}.$$

The choices in (15) make this at most ε. In both cases the mass is at most $\alpha e^{\kappa} \le (1+\varepsilon)\alpha$.

The bisection bracket and its final gap. Maintain certified endpoints $c_- < c_* < c_+$, each with a stored boundary candidate. Midpoints use the same three-way sign test. Estimated objective values $\widehat{F}_-, \widehat{F}_+$ have error at most ν and their candidates have objective error at most η. Thus

$$m(c_-) - m(c_+) \le \widehat{F}_- - \widehat{F}_+ + 2\nu + \eta =: U. \tag{26}$$

When $c_+ - c_- \le q$ and $U \le q$, return the right candidate w_+. The target's exact gap obeys $\Delta = m(c_*) - m(c_+) \le q$, so its KL error is at most $\eta + D\sqrt{2M_S q} \le \varepsilon$. Moreover,

$$\mathcal{L}(w_+) - \mathcal{L}(a) \le c_+ - c_* \le q, \qquad \log[p_{w_+}(S)/\alpha] \le \eta + \zeta,$$

where the latter follows from $r(c_+) \le \alpha$ and (24). This proves the remaining two guarantees. Both the bracket width and the objective gap must meet their stopping bounds.

For termination, Lemma 14 and Lemma 5 give on the whole bracket region

$$\|\nabla F_S\| \le B := L\sqrt{2H/\lambda} + CK\ell.$$

Equation (21) implies that $m(c_-) - m(c_+) \le q/2$ once the bracket width is at most $\lambda q^2/(8B^2)$. The definition of U also gives $U \le m(c_-) - m(c_+) + 2\eta + 4\nu \le 7q/8$. Thus at most

$$O\left(1 + \log_+ \frac{g}{\min\{q, \lambda q^2/(8B^2)\}}\right)$$

bisection calls suffice, unless a candidate with mass close to α is returned sooner; $\log_+ t = \max\{0, \log t\}$. Together with $O(1 + \ell/g)$ grid calls this provides a deterministic bound Q. The cost of finding a candidate near the boundary in Appendix D and the mass-estimation cost (23) give the polynomial bounds in Theorem 8. The observations enter through their empirical sufficient-statistic mean and NLL; the data need not follow a conditional model in the family.

G Hardness without the capacity assumption

Proof of Claim 1. Reduce from the decision version of Max-Cut (Karp, 1972). Its input is a graph $G = (W, E)$ with $d \ge 2$ vertices, nonnegative integer edge weights w_{ij}, and an integer threshold b. For $s \in \{-1, 1\}^d$, let

$$\mathrm{cut}_G(s) = \sum_{\{i,j\} \in E} w_{ij} \mathbf{1}_{\{s_i \ne s_j\}} = \frac{1}{4} \sum_{\{i,j\} \in E} w_{ij}(s_i - s_j)^2.$$

Put $R = 8d$, $r = 2d$, and define the valid set

$$V = \{0\} \cup \bigcup_{s:\,\mathrm{cut}_G(s)\geq b} \{x : \|x - Rs\|_\infty \leq r\}, \qquad S = \mathbb{R}^d \setminus V. \tag{27}$$

Although the union may contain exponentially many boxes, membership is computable in polynomial time: except for 0, a point lies in at most one box, whose center's sign vector is obtained from its coordinate signs. Check its distance to that center and evaluate the cut. All constants have polynomial bit complexity. Take the specified valid observation to be $x_1 = 0$; the isolated point has zero Gaussian probability and does not affect the feasibility constraint.

If the graph has a cut s of size at least b, choose $\theta = Rs$. For $Z \sim \mathcal{N}(\theta, I_d)$, Markov's inequality gives

$$\Pr(\|Z - \theta\|_\infty > 2d) \leq \Pr(\|Z - \theta\|_2^2 > 4d^2) \leq \frac{1}{4d} \leq \frac{1}{4}.$$

Thus $p_\theta(V) \geq 3/4$, yielding a feasible parameter. If there is no such cut, $V = \{0\}$ has zero Gaussian mass for every θ, and no feasible parameter exists. This proves NP-hardness of the feasibility decision with a constant gap.

One can also recover a cut from any feasible returned parameter. If $p_\theta(V) \geq 3/4$, the event $\|Z - \theta\|_\infty \leq 2\sqrt{d}$ has probability at least $3/4$, again by Markov. It intersects $Z \in V \setminus \{0\}$ with positive probability. Hence for some valid box center Rs,

$$\|\theta - Rs\|_\infty \leq 2d + 2\sqrt{d} \leq 4d < R,$$

so the coordinate signs of θ equal s, and its cut meets the threshold. The result concerns arbitrary efficiently testable validity sets. It does not contradict the positive theorem, which additionally requires powerfulness. □

H GAUSSIAN ILLUSTRATION

A synthetic one-dimensional Gaussian illustration shows the tradeoff between likelihood and invalid mass. The validity set depends on the observed data. The experiment does not test population generalization or establish all assumptions of Theorem 8.

The experiment uses $n = 40$ valid observations: right examples from $\mathcal{N}(1, 2^2)$ conditioned on $x \geq 0$ and left examples from $\mathcal{N}(-1, 2^2)$ conditioned on $x < 0$. These labels are only for visualization. Given the negative observations x_i^-, the valid and invalid sets are

$$V = \{x \geq 0\} \cup \bigcup_i [x_i^- - d_0, x_i^- + d_0], \qquad S = \mathbb{R} \setminus V,$$

where $d_0 > 0$ controls the valid islands. The ordinary maximum-likelihood fit ignores S. The illustrated update uses likelihood sublevels and invalidity queries to reduce mass on S.

Across ten random seeds, the initial hallucination rate is 0.5239 ± 0.0151 and the final rate is 0.0338 ± 0.0038, with errors denoting standard errors of the mean. All ten runs fall below the target $\alpha = 0.05$. The conditional fit to valid observations, $\mathcal{L} + \log p_\theta(V)$, is distinct from the optimized invalid-set objective $F_S = \mathcal{L} + \log p_\theta(S)$; no numerical claim about the latter is inferred from a valid-conditional loss.

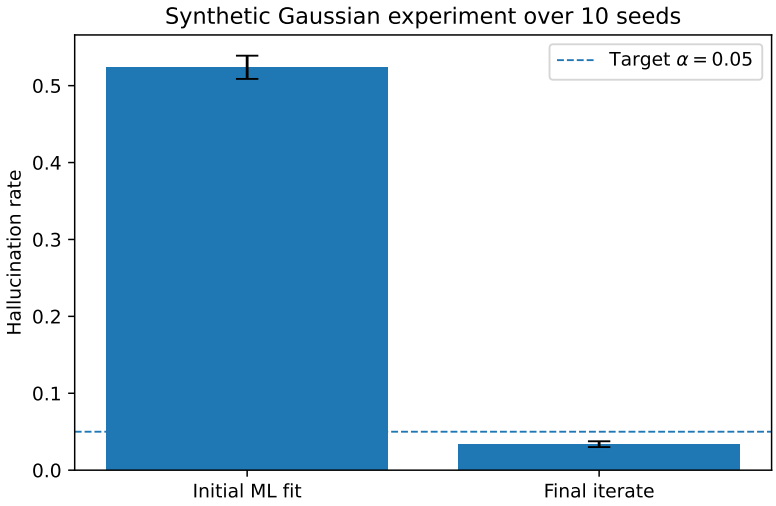


Figure 2: Hallucination rates over ten random seeds. Bars show the mean for the initial and final fits, with standard errors. The dashed line is the target $\alpha = 0.05$.

I Log-partition identities and conditional KL

We record the exponential-family calculations used in the main text. They apply to the fixed family and do not require the data distribution to be a member of that family.

Lemma 17 (Differentiation of the restricted log partition)**.** *Let Θ be an open natural-parameter domain on which A is finite, and let S have positive probability under the common support. Then A_S is twice differentiable on Θ and*

$$\nabla A_S(\theta) = \mathbb{E}_{p_\theta^S} T, \qquad \nabla^2 A_S(\theta) = \mathrm{Cov}_{p_\theta^S}(T).$$

Consequently $F_S = A_S - \theta^\top \widehat{\mu} + C$ is convex and

$$\nabla \log p_\theta(S) = \mathbb{E}_{p_\theta^S} T - \mathbb{E}_{p_\theta} T, \qquad \mathrm{KL}(p_a^S \| p_b^S) = F_S(b) - F_S(a) - \nabla F_S(a)^\top (b - a).$$

Proof. Fix $\theta \in \Theta$ and a closed coordinate box around it contained in Θ. On a smaller box, any polynomial of degree at most two in the coordinates of T, multiplied by $e^{\vartheta^\top T}$, is bounded by a constant times the sum of $e^{\eta^\top T}$ over the vertices η of the larger box. Each term is integrable against h because $A(\eta) < \infty$. This supplies a common integrable bound for the first two derivatives, including after multiplication by $\mathbf{1}_S$.

Write $Z_S(\theta) = \int_S h(x) e^{\theta^\top T(x)} \, \mathrm{d}x$. Differentiation under the integral gives $\nabla Z_S = \int_S T h e^{\theta^\top T}$ and $\nabla^2 Z_S = \int_S T T^\top h e^{\theta^\top T}$. The quotient rule for $A_S = \log Z_S$ yields the conditional mean and covariance. A covariance matrix is positive semidefinite, proving convexity. Since $\log p_\theta(S) = A_S(\theta) - A(\theta)$, subtraction gives the mass-gradient identity.

Finally, on the common conditional support,

$$\log \frac{p_a^S(x)}{p_b^S(x)} = (a - b)^\top T(x) - A_S(a) + A_S(b).$$

Its expectation under p_a^S is finite by the preceding derivative bounds. Substituting $\mathbb{E}_{p_a^S} T = \nabla A_S(a)$ gives the displayed KL formula for A_S. The affine term $-\theta^\top \widehat{\mu} + C$ cancels, giving the same identity for F_S. □

J Convex calculations for nested sublevels

The following proof gives the version of the nested-progress argument needed here; compare Karatapanis et al. (2025, Proposition C.1). Only convexity and attainment of the sublevel minima are used.

Proof of Proposition 7. Take $0 \le c < e \le H$, minimizers $x_c \in C_c$ and $x_e \in C_e$, and $t \in [0,1]$. Convexity of $\mathcal{L}$ gives

$$\mathcal{L}((1-t)x_c + tx_e) \le (1-t)(\mathcal{L}_0 + c) + t(\mathcal{L}_0 + e),$$

so this point belongs to $C_{(1-t)c+te}$. Convexity of F_S then implies

$$m((1-t)c + te) \le F_S((1-t)x_c + tx_e) \le (1-t)m(c) + tm(e).$$

This is convexity of m. The inclusions $C_c \subseteq C_d$ for $c < d$ show that m is non-increasing. Set $t = (d-c)/(e-c)$ in the preceding inequality and rearrange to obtain (9), including $d = e$.

For a grid spacing g and $c + g \le e$, the inequality becomes

$$m(c) - m(e) \le \frac{e-c}{g}\,[m(c) - m(c+g)].$$

To account for approximation, suppose candidates at c and $c+g$ have objective errors in $[0,\eta]$, and their objective evaluations have absolute error at most ν. If the difference of the two estimated objective values is at most τ, then

$$m(c) - m(c+g) \le \tau + \eta + 2\nu.$$

Indeed, the first estimated value is at least $m(c) - \nu$, while the second is at most $m(c+g) + \eta + \nu$. The resulting bound on later improvement is at most $(H/g)(\tau + \eta + 2\nu)$. It controls an objective gap; returning a parameter still requires the mass check described in the algorithm. □

Proof of Observation 6. For a boundary minimizer b_c and any $\theta \in C_c$, the inequality $F_S(\theta) \ge F_S(b_c)$ gives

$$\log p_\theta(S) \ge \log p_{b_c}(S) + \mathcal{L}(b_c) - \mathcal{L}(\theta).$$

Exponentiating proves the first claim. Since m is non-increasing,

$$\log r(d) - \log r(c) = m(d) - m(c) - (d-c) \le -(d-c),$$

which proves the second. Thus r is strictly decreasing even when m is locally constant. Together with the continuity bound in Lemma 14, this justifies the comparisons with the unique target threshold c_*. □

K Sampling estimates and repeated optimization

K.1 Estimating invalid mass

Lemma 18 (Log-mass estimation and the initial threshold test)**.** *Let $B_1, \dots, B_N$ be independent Bernoulli variables with mean p, and let $\widehat{p} = N^{-1}\sum_i B_i$. For $0 < \alpha < 1$, $c_0 > 0$, $0 < \nu < 1/2$, and $0 < \delta' < 1/2$, if $p \ge c_0\alpha$, then*

$$N \ge \frac{C}{c_0\alpha\nu^2}\log\frac{2}{\delta'} \quad\Longrightarrow\quad |\log\widehat{p} - \log p| \le \nu$$

with probability at least $1-\delta'$. With no lower bound on p, $N \ge C\alpha^{-1}\nu^{-2}\log(2/\delta')$ suffices for the following test: return the initial parameter if $\widehat{p} \le \alpha e^{2\nu}$; otherwise continue using $\log\widehat{p}$. With probability at least $1-\delta'$, a returned parameter has $p \le \alpha e^{4\nu}$, and continuation implies $p > \alpha$ and log-estimation error at most ν.

Proof. For any real t, independence and $1 + x \le e^x$ imply

$$\mathbb{E}e^{t\sum_i B_i} = (1 - p + pe^t)^N \le \exp[Np(e^t - 1)].$$

The exponential Markov inequality, optimized at $t = \log(1 + r)$ for the upper tail and $t = \log(1 - r)$ for the lower tail, gives

$$\Pr(|\widehat{p} - p| > rp) \le 2e^{-Npr^2/3}, \qquad 0 < r < 1.$$

The scalar inequalities used here are $(1+r)\log(1+r) - r \ge r^2/3$ and $r + (1-r)\log(1-r) \ge r^2/2$; both follow by differentiating on $[0, 1)$. Set $r = \nu/2$. On the resulting event, $|\log(\widehat{p}/p)| \le 2r = \nu$, which proves the first assertion and also ensures $\widehat{p} > 0$.

For the initial test, if $p \ge \alpha/2$, apply the same bound with this lower mass bound. A returned parameter then satisfies $p \le \widehat{p}e^{\nu} \le \alpha e^{3\nu}$, whereas continuation gives $p \ge \widehat{p}e^{-\nu} > \alpha e^{\nu}$. If $p < \alpha/2$, couple the observations with Bernoulli variables of mean $\alpha/2$ that dominate them. Their upper-tail bound at $t = \log 2$ shows $\Pr(\widehat{p} > \alpha) \le e^{-cN\alpha}$. Except on this event the test returns, and $p < \alpha$ already. Adjusting C proves the stated bounds for all p, including $p = 0$ without taking its logarithm. □

K.2 Rejection sampling and high-probability costs

An accepted proposal from p_θ, selected by $\mathbf{1}_S = 1$, has law p_θ^S: for measurable E, its probability is $p_\theta(E \cap S)/p_\theta(S)$. The waiting time J has $\Pr(J > m) = (1 - p_\theta(S))^m$ and expectation $1/p_\theta(S)$. On a region with mass at least $\beta = \alpha/e$,

$$\Pr(J > m) \le e^{-\beta m}.$$

For at most M adaptive conditional calls, cap each call at $\lceil \beta^{-1} \log(M/\delta') \rceil$ proposals and declare failure if that cap is exceeded. Conditional on the preceding history, the same tail bound applies to each call. A union bound therefore gives total cost at most $M\lceil \beta^{-1} \log(M/\delta') \rceil$ with failure probability at most δ'. Couple capped and uncapped runs using the same proposals. They agree unless a cap is exceeded, so the capped algorithm differs from the ideal oracle implementation with probability at most δ'. This makes the additional polynomial dependence on $1/\alpha$ explicit.

K.3 Selecting a successful optimization run

Here are the details behind the repetition step in Appendix D. Let $p_0 > 0$ be a lower bound on the probability that a random direction has the alignment required by Proposition 9. The PSGD second-moment bound is uniform in the unit direction. Choose N so that, conditional on every direction, the expected proxy error is at most $\xi/4$. Markov's inequality gives proxy error at most ξ with conditional probability at least $3/4$. Hence a single run has both suitable alignment and proxy accuracy with probability at least $3p_0/4$.

With the tolerances of Appendix D, such a run has objective error at most $\eta/4$ and boundary slack at most $\zeta/2$. For independent runs, the probability that no such candidate occurs is at most $(1 - 3p_0/4)^R \le e^{-3p_0R/4}$. Taking $R \ge (4/(3p_0)) \log(Q/\delta')$ controls all Q sublevel solves by a union bound. Suppose objective estimates have error at most $\eta/8$. Among the candidates passing the boundary test, the one with smallest estimated objective has actual objective at most that of a successful candidate plus $\eta/4$, hence at most $m(c) + \eta/2$. The remaining margin permits the numerical errors stated earlier. A failed individual run does not spoil the comparison, because selection is based on objective values rather than on identifying which direction succeeded.

K.4 Effect of numerical projection and gradient bias

The elementary PSGD calculation also quantifies these errors. Suppose all iterates are in a convex set C of diameter $D > 0$, the conditional gradient bias has norm at most b, and the conditional second moment is at most G_f^2. If the next iterate is feasible and within distance r of the exact projection of $w_t - hg_t$, then

$$\mathbb{E}[f(\overline{w}) - \min_C f] \le \frac{D^2}{2hN} + \frac{hG_f^2}{2} + Db + \frac{2Dr + r^2}{2h}.$$

To prove this, let $y = \Pi_C(w_t - hg_t)$ and $w^* \in \arg\min_C f$. Since $y, w^* \in C$ and $\|w_{t+1} - y\| \le r$,

$$\|w_{t+1} - w^*\|^2 \le \|y - w^*\|^2 + 2Dr + r^2 \le \|w_t - hg_t - w^*\|^2 + 2Dr + r^2.$$

After expansion and conditional expectation, the bias contributes at most $2hDb$. Summing and using convexity at the averaged iterate gives the displayed bound. For a target error ξ, choose the first two terms to sum to at most $\xi/2$, take $b \le \xi/(4D)$, and take $r \le \min\{h\xi/(8D), \sqrt{h\xi/4}\}$. The remaining terms sum to at most $\xi/2$. A singleton sublevel needs no optimization. Thus feasible approximate projections can be assigned explicit inverse-polynomial tolerances; the sampling assumptions remain valid at every iterate.