

Contrastive guidance

Conditioning, exponential tilting and flow policies

This is an AI-assisted explanatory derivation based on Advantage Guidance v2, not a new theorem attributed to the paper or a benchmark reproduction. The Markdown note is canonical. Exact authoring model and effort are not exposed by the authoring run; the owner has reviewed the notes and companion outputs. This is not an independent scientific audit.

Navigate to [the variational optimum](#), [the worked example](#), or [the score–velocity relationship](#). Equations are numbered for reference.

Scope and notation

This is a derivation notebook about **conditioning, exponential tilting and generative policy extraction**. Its main paper is [Advantage Guidance v2, Sections 3–6 and Appendix A.1](#). The mathematical steps below are worked explanations; the finite discrete example and tests are notebook constructions, not benchmark results. It begins with one paper and can grow into a multi-paper concept note as further sources are actually read.

Fix a state s and suppress it when convenient. Write $q(a)$ for a reference density, $A(a)$ for a fixed estimated advantage, and $\beta > 0$ for a regularization temperature. Probability statements concern this fixed state; they do not solve the entire sequential control problem. Assume the densities and derivatives used below exist, and that the relevant normalizing integrals are finite.

1. Derive the KL-regularized optimum

Consider policies absolutely continuous with respect to q and maximize

$$J(p) = \int p(a) A(a) da - \beta \int p(a) \log \frac{p(a)}{q(a)} da, \quad (1)$$

subject to $\int p(a) da = 1$. With a multiplier λ , the stationary condition is

$$A(a) - \beta \left[\log \frac{p(a)}{q(a)} + 1 \right] + \lambda = 0. \quad (2)$$

Rearrangement gives $p(a) = q(a)e^{A(a)/\beta} C$, with C determined by normalization. Therefore

$$p^*(a) = \frac{q(a)e^{A(a)/\beta}}{Z}, \quad Z = \int q(a)e^{A(a)/\beta} da. \quad (3)$$

A stronger check substitutes this density back into the objective:

$$J(p) = \beta \log Z - \beta D_{\text{KL}}(p \parallel p^*). \quad (4)$$

Nonnegativity of KL proves the optimum, rather than relying only on a stationary point. This requires $Z < \infty$; an unrestricted unbounded utility need not satisfy that condition. Where q is zero, this objective

cannot create new support at finite KL. In practical neural policies, support may be formally broad while reliable offline-data coverage is narrow; those are not the same assurance.

At fixed state, adding a constant $c(s)$ to A changes Z by the same multiplicative factor and leaves p^* unchanged. Multiplying advantage by a positive constant changes the effective temperature. These are useful invariance checks for both derivations and implementations.

The result is optimal for the **specified fixed-state surrogate objective**. It does not show that the critic is accurate, that the state visitation distribution is unchanged after deployment, or that the resulting policy globally optimizes an MDP.

2. Bayes conditioning creates a matching preference

Let a noisy scalar label satisfy

$$Y \mid a \sim \mathcal{N}(A(a), \sigma_y^2), \quad \sigma_y > 0. \quad (5)$$

Bayes' rule yields

$$p(a \mid y) = \frac{q(a)p(y \mid a)}{p(y)} \propto q(a) \exp \left[-\frac{(y - A(a))^2}{2\sigma_y^2} \right]. \quad (6)$$

The label y is a value to match. Its likelihood is largest when $A(a) = y$ and decreases on either side. For an action with advantage already greater than y , improving the advantage further makes this likelihood worse. That is the mismatch between conditioning and optimization.

This is a model of the clean action-label joint distribution. In AG's implementation, labels are based on estimated, state-normalized advantages, not an oracle utility. Every use of the ideal identity should specify which function occupies the symbol A .

3. What ordinary CFG changes

Let $g = \nabla_a \log q(a)$ and $g_y = \nabla_a \log p(a \mid y)$. Because $p(y)$ is constant with respect to a ,

$$g_y - g = \nabla_a \log p(y \mid a) = \frac{y - A(a)}{\sigma_y^2} \nabla_a A(a). \quad (7)$$

Then the CFG combination is

$$g + w(g_y - g) = \nabla_a \log \left[q(a) e^{-w(y - A(a))^2 / (2\sigma_y^2)} \right]. \quad (8)$$

For $w > 0$, the direction due to the label changes sign when A crosses y . Larger w strengthens the same matching tendency. At $w = 1$, the clean-density expression is simply the conditional model. This does not imply that every flow implementation with this numerical setting yields exact conditional samples: the statement here is about the assumed density and its score.

A request above all achievable advantages can favor the best available actions in a toy bounded setting. But choosing such a request requires knowing the range, and a learned conditional model may be unreliable there. That workaround does not establish equality with exponential tilting at an arbitrary temperature.

4. Why the contrastive ratio cancels the quadratic term

For two labels, subtract their log likelihoods:

$$\log \frac{p(y_+ \mid a)}{p(y_- \mid a)} = \frac{(y_- - A)^2 - (y_+ - A)^2}{2\sigma_y^2}. \quad (9)$$

Expanding both squares cancels A^2 :

$$\log \frac{p(y_+ | a)}{p(y_- | a)} = \frac{y_+ - y_-}{\sigma_y^2} A(a) + C, \quad (10)$$

where C does not depend on action. The same action-dependent term appears in the ratio of conditional action densities because their reference density cancels and their normalizers are action-independent. Thus

$$g_{y_+} - g_{y_-} = \frac{y_+ - y_-}{\sigma_y^2} \nabla_a A(a). \quad (11)$$

Adding the difference to the reference score gives

$$g + w(g_{y_+} - g_{y_-}) = \nabla_a \log[q(a)e^{cA(a)}], \quad c = \frac{w(y_+ - y_-)}{\sigma_y^2}. \quad (12)$$

For $c > 0$, this is the target score with $\beta = 1/c$. Reversing the labels reverses the preference; identical labels remove the guidance. Those are simple checks that catch sign errors.

The exponential-family generalization

Suppose instead

$$p(y | A) = h(y) \exp\{T(y)A - F(A)\}. \quad (13)$$

The difference of log likelihoods cancels $F(A)$, leaving $[T(y_+) - T(y_-)]A$ plus a constant. Consequently $1/\beta = w[T(y_+) - T(y_-)]$. This is the structure used in Section 6.1 and Appendix A.1. It requires A to occupy the specified natural-parameter position; merely saying that a distribution “is exponential-family” is insufficient.

For Gaussian labels with constant variance, $T(y) = y/\sigma_y^2$. If variance itself depends on action, new action-dependent terms appear and the simple cancellation no longer applies. Likewise, zero label noise is a singular limit of this density derivation, even if a zero-noise empirical variant happens to work.

5. A worked discrete example

Take three equally likely actions with utilities $A = [0, 1, 2]$. Let $\sigma_y = 1$, $w = 1$, and compare ordinary conditioning at $y = 1$ with the contrastive pair $(y_+, y_-) = (1, 0)$.

Utility	Reference	Gaussian conditioning / CFG at $w=1$	Contrastive target, $\beta=1$
0	0.333333	0.274069	0.090031
1	0.333333	0.451863	0.244728
2	0.333333	0.274069	0.665241

Conditioning gives utilities zero and two equal probability because both are one unit from the query. Contrastive weighting prefers two, then one, then zero. Expected utility is 1 for the matching distribution and approximately 1.575210 for the tilted distribution.

The computation is just normalized weights: $\exp[-(1 - A)^2/2]$ versus $\exp(A)$. It requires no trained model and no random sampling. The repository tests reconstruct the contrastive likelihood ratio and compare it with the direct tilt. They also verify the KL objective decomposition on several candidate distributions. Passing these tests verifies arithmetic and algebra in this example; it does not validate the paper’s benchmark or a neural sampler.

6. From score space to straight-path flow velocity

It is tempting to replace every score in the derivation with a velocity and call the result exact. Here is the extra structure that needs to be stated.

Let a clean action be a , independent standard Gaussian noise be ϵ , and $x_t = ta + (1 - t)\epsilon$. For $0 < t < 1$, write $g_t(x) = \nabla_x \log p_t(x)$. Gaussian denoising gives

$$g_t(x) = -\frac{x - t\mathbb{E}[a | x]}{(1 - t)^2}. \quad (14)$$

The population flow-matching velocity is

$$v_t(x) = \mathbb{E}[a - \epsilon | x] = \frac{\mathbb{E}[a | x] - x}{1 - t}. \quad (15)$$

Eliminating the posterior mean gives

$$v_t(x) = \frac{x}{t} + \frac{1 - t}{t}g_t(x). \quad (16)$$

The same relation holds conditionally on a label when the interpolation and noise law are shared. Hence conditional velocity *differences* are proportional to conditional score differences at that intermediate time; the x/t term cancels. These relations are derived here for this particular path, not asserted for arbitrary noise schedules. Endpoint expressions require limits; evaluating x/t at $t = 0$ is invalid, even though a learned velocity can be evaluated there.

Now the critical limitation: the Gaussian label model was specified given the **clean action**. Given an intermediate x_t , the label likelihood averages over possible clean actions. That mixture generally is not a Gaussian with mean $A(x_t)$ and the same variance. The clean-action cancellation therefore does not automatically prove the same tilted marginal identity at every intermediate time.

Even a score assembled from an instantaneous product of marginals need not describe a consistent forward-noising path of the desired final density. The practical guided ODE must still be evaluated. In addition, learned velocity errors and Euler discretization introduce further error. These are separate from a sign mistake in the clean-density algebra.

7. Residual parameterization in that sampler

With $v_y = v_{\text{ref}} + v_\theta(y)$, the practical rule is

$$v_{\text{guided}} = v_{\text{ref}} + w[v_\theta(y_+) - v_\theta(y_-)]. \quad (17)$$

The shared reference cancels inside the difference. Algorithm 1 integrates this velocity from Gaussian noise toward an action. The residual networks must be queried at identical state, time and current noisy action. Querying the two labels from independently sampled intermediate actions would no longer be a pointwise difference of fields.

Residual training and action imagination help the model learn label-dependent fields. They do not remove the need to learn a critic. A useful implementation review separately checks reference-actor updates, critic targets, advantage normalization, residual loss gradients and sampler field evaluations rather than judging correctness from one end-to-end reward curve.

8. Normalization changes the units of guidance

Suppose $\tilde{A} = (A - \mu_s)/d_s$, where $d_s = \sqrt{\text{var}_s(A) + \delta} > 0$. At fixed state and treating these statistics as fixed,

$$q(a)e^{c\tilde{A}(a)} \propto q(a)e^{(c/d_s)A(a)}. \quad (18)$$

The subtraction of μ_s cancels from normalization; the division does not. The effective raw-advantage temperature is $\beta_{\text{raw}} = d_s/c$. States with different scales receive different raw-unit preferences even when w , label spacing and label noise are held fixed.

For example, let $A = [2, 4, 6]$, $\mu = 4$, $d = 2$, and $c = 1$. Standardized utilities are $[-1, 0, 1]$. Their tilted probabilities equal those obtained from raw utility at $\beta = 2$: approximately $[0.090031, 0.244728, 0.665241]$. They do not equal the raw tilt at $\beta = 1$.

AG estimates its normalization statistics from the dataset action and sampled alternatives. Finite-sample variability, tiny within-state variance and the stabilizer all matter. The fixed-statistics derivation is an interpretation of a particular estimated normalization, not a claim that these statistics are globally exact or independent of the proposal policy.

9. What is established, and what remains to test

There are four levels of evidence to keep separate:

1. **Variational identity:** the exponential tilt solves the stated fixed-state KL objective.
2. **Clean-score identity:** the contrastive ratio produces that score under the assumed likelihood model.
3. **Learned sampler:** neural approximation, intermediate-time guidance and numerical integration determine actual generated actions.
4. **Control performance:** the critic and environment determine whether those actions improve real return and satisfy constraints.

The first two do not certify the last two. A good next experiment holds a known utility fixed, varies integration steps and learned-field accuracy, and compares the generated distribution against an exactly normalized small reference problem. A second experiment perturbs the utility estimator while evaluating with the unperturbed utility. This separates sampler error from critic exploitation.

For future literature connections, add a separately read source on flow matching or classifier-free guidance and state precisely which path, marginal or endpoint theorem it provides. Until then, this is a worked technical companion to one paper, not a claim to cover the whole literature.

Note provenance

Authoring tool: Codex. Model: not exposed by this run. Reasoning setting: not exposed by this run. Generated/updated: 2026-10-30.

Review: same-session source checks and deterministic mathematical tests. Review model: not exposed by this run. Review date: 2026-10-30. Independent review: no.

Human review: The owner reviewed the notes and companion outputs. This is not an independent scientific audit. Review date: 2026-10-30.